Bioinformatics 2019-2020
Date: 03 January 2020
Project report:

Bioinformatics@Data Science A.Y. 2019-2020

Network Medicine Project: Human Primary Microcephaly (MCPH) genes analysis

Ivana Nastasic and Joanna Broniarek

Group no. 14

Abstract

Primary Microcephaly (MCPH) is a disorder manifested as a reduction in head circumference and this clinical finding infers that an individual has a significant diminution in brain volume. In Primary Microcephaly the brain fails to grow to the correct size during pregnancy. The analysis show that the Fanconi anemia pathway that is required for the efficient repair of damaged DNA is crutial for the disease. Mutations of analysed seed genes could have an impact, for instance on the mitotic spindle orientation or on the signalling responses, because of the damaged DNA, and further leading to the primary microcephaly disease. Among the putative disease genes modules obtained through the whole analysis, some of the genes from MCPH1-MCPH12 loci were also found.

1. Introduction

Primary microcephaly (MCPH) is a disorder of brain development characterised by a reduced occipitofrontal circumference of the head (at least 3 standard deviations below the mean for age and gender). It has a wide variety of causes, including toxic exposures, in utero infections, and metabolic conditions [2]. MCPH exhibits genetic heterogeneity, and to date, twelve loci (MCPH1-MCPH12) have been associated with this condition and they contain the following genes: Microcephalin, WDR62, CDK5RAP2, CASC5, ASPM, CENPJ, STIL, CEP135, CEP152, ZNF335, PHC1 and CDK6. It is predicted that MCPH gene mutations may lead to the disease phenotype due to a disturbed mitotic spindle orientation, premature chromosomal condensation, signalling response as a result of damaged DNA, microtubule dynamics or a few other hidden centrosomal mechanisms that can regulate the number of neurons produced [3]. An in-depth analysis of the genes associated with primary microcephaly (seed genes) and their protein-protein interactions could help to better understand the mechanisms of this disease.

2. Seed genes

We explored the DisGeNET dataset in order to collect the list of genes associated to primary microcephaly. DisGeNET is one of the largest available collections of gene-disease associations [4].

We used the Curated Gene Disease Associations dataset from DisGeNet and collected human genes involved in the Primary microcephaly (DisGeNet ID:C0431350). As a result we obtained a list of 109 genes.

Then, we mapped the list of genes with Entrez gene IDs to the Uniprot website [6] in order to gather some basic information like: primary gene symbol, Uniprot AC, protein name and description of its function. We noticed that gene RNU4ATAC (U4atac small nuclear RNA) was not mapped and corresponds to small nuclear RNA, so we excluded it from further protein-protein interactions analysis. Moreover, we found a gene ERCC6

encoding proteins with two different UniprotAC: Q03468 and P0DP91. We checked that the second protein (P0DP91) does not have any entry associated to the Biogrid database, hence we decided to keep only the first one. In addition, we checked if all the collected gene symbols are approved on the HGNC website. We found seven genes which did not match the HGNC approved gene symbols, so we corrected them. Table 1 contains the mentioned not approved genes.

Table 1: **Seed genes with gene symbols not approved by HGNC symbols.** For the analysis, genes from the Collected Gene Symbols column were converted into associated Approved Symbols.

Collected Gene Symbol	Match type	Approved Symbol
IARS	Previous symbol	IARS1
NBN	Alias symbol	ARTN
ORC1	Alias symbol	SLC25A15
QARS	Previous symbol	EPRS1
QARS	Previous symbol	QARS1
RPL10	Alias symbol	RPL15
BRIP1	Alias symbol	MRPL36

Finally, we obtained 108 seed genes included in Table 12. Because of the length, we located the table at Appendix.

3. Interaction data collection

For each seed gene, we collected all the binary protein interactions from two different protein-protein interaction (PPI) sources:

- Biogrid Human, the latest release available
- IID Integrated Interactions Database

BioGRID

We checked with the use of Python code if all interactors gathered (A and B) are from Human. We found some interactions with proteins from different organisms, and excluded all the interactions associated with them. Additionally, we limited the results only to physical interactions.

As the next step, we uploaded the BioGrid interactions file to the Cytoscape software, together with the seed genes table. Then, we retrieved a list of the proteins interacting with at least one seed gene, including also interactions among the non-seed proteins in our network. From described network, we extracted the edge table and parsed it with the use of Python.

We used the Uniprot mapping online tool to match and gather UniprotAC and Gene Symbol for each non-seed gene ID. The noticed some missing or duplicated records. Hence, we analysed and preprocessed obtained results in a following way:

- among unmapped genes there were 72 miRNA, that we excluded from the interactome
- excluded genes without mapped Gene Symbol, i.e. geneID : 101929876. These records represented genes that are not currently annotated and have been withdrawn by NCBI staff.
- there were 11 genes with multiple Uniprot AC. We manually checked and excluded not necessary Uniprot AC (i.e. they had the same Biogrid association or one of them did not have any). Table 2 presents the selected UniprotAC for such genes.

Table 2: **Gene Symbols with selected UniprotAC associations from the BioGrid Interactome.** For genes with more than one UniprotAC, the selection of UniprotAC was done manually, taking into account if they have a reference to the Biogrid database.

Gene Symbol	Uniprot AC
AKAP7	O43687
BBC3	Q9BXH1
CDKN2A	P42771
CUX1	P39880
ERCC6	P0DP91
MOCS2	O96007
RAB34	P0DI83
RABGAP1L	Q5R372
TMPO	P42166
TOR1AIP2	Q8NFQ8
ZNF365	Q70YC4

The summary of main results from the Biogrid interactions are included in the Table 3.

Integrated Interactions Database (IID)

In order to obtain all required interactions from Integrated Interactions Database (IID), as the first step we downloaded the protein-protein interaction for seed genes. We filtered there results for experimental data only. Then, we collected all the interacting genes (also non-seed genes) and uploaded them once again to the IID database. As a result, we gathered the interactions with seed genes as well as with non-seed genes.

With the use of Cytoscape software, we extracted all the required interactions with seed-genes, between seed genes and between non-seed genes (only first neighbours). To be sure that all interacting proteins come from Human, we checked it with Uniprot website. The summary of interactions from the IID database are included in the Table 3.

Table 3: **Summary of the Protein-Protein-Interactions for BioGrid and IID database.**

Database	Number of found seed genes	Number of interacting proteins (seed genes included)	Number of interactions
BIOGRID	105	4716	236512
IID	106	4075	124764

From previously collected and cleaned interactions using Python, we built following datasets:

- **Seed Genes Interactome (SGI)** - contains interactions that involve seed genes only, from all DBs.
- **Union Interactome (U)** - contains proteins interacting with at least one seed gene, from all DBs.
- **Intersection Interactome (I)** - includes all proteins interacting with at least one seed gene confirmed by both DBs.

4. Enrichment analysis

In order to analyse if selected genes, based on the current scientific knowledge, are particularly associated to some biological processes and functions, we used the services provided by EnrichR [7]. We performed GO (gene ontology) analysis for Biological Process, Molecular Function and Cellular Component, and Pathway analysis with KEGG. These two analysis were ran over the Seed Genes Interactome (SGI) and the Union Interactome (U). The first 10 most significant results, based on p-value are presented in Chart 1-4 for SGI and Chart 5-8 for U.

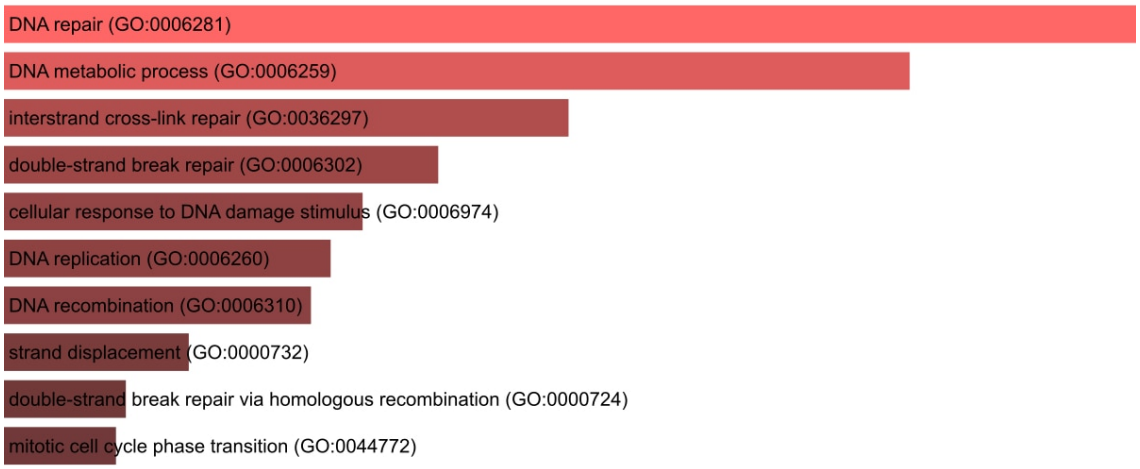

Figure 1. GO Biological Process - Seed Genes Interactome. The first 10 most significantly enriched GO Biological Processes by genes involved in Seed Gene Interactome. The color of the bars corresponds to the p-value (the ligher, the more significant).

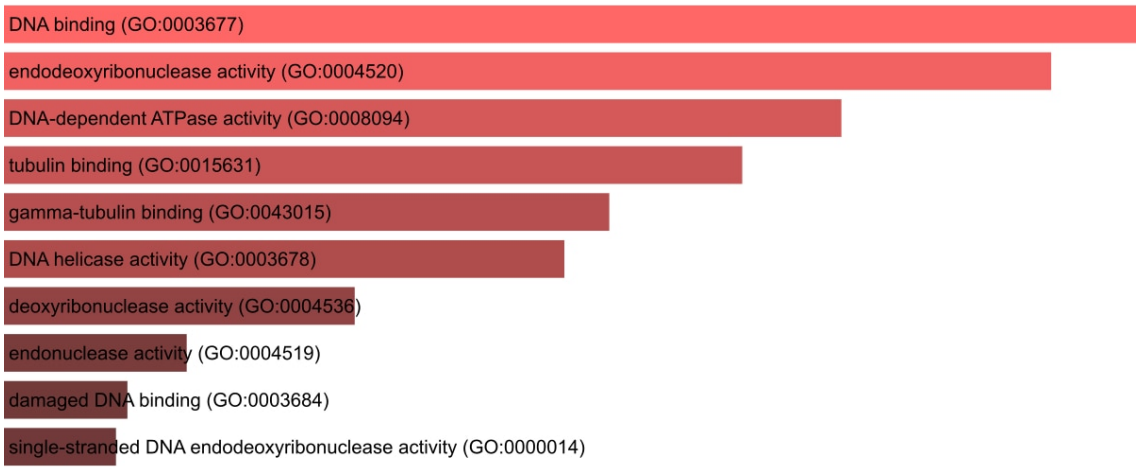

Figure 2. GO Molecular Function - Seed Genes Interactome. The first 10 most significantly enriched GO Molecular Function by genes involved in Seed Gene Interactome. The color of the bars corresponds to the p-value (the ligher, the more significant).

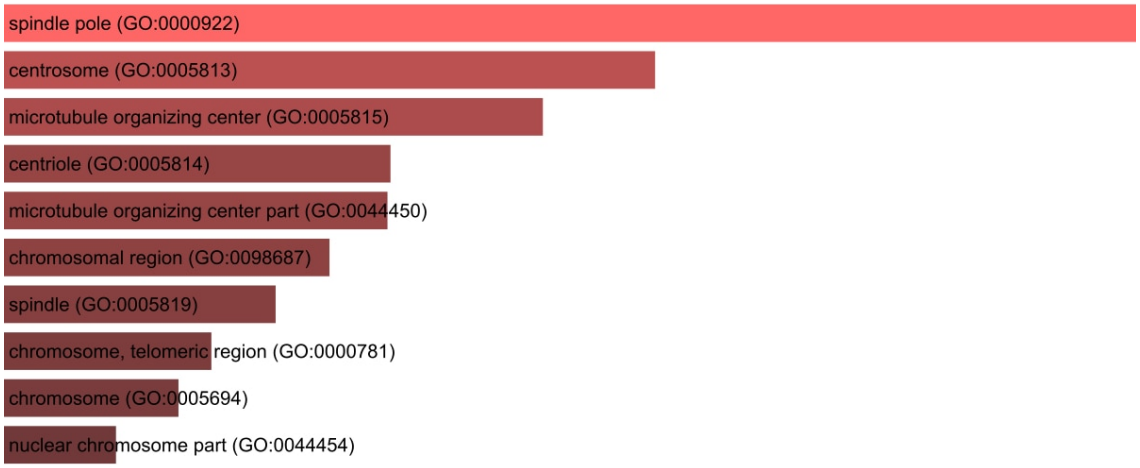

Figure 3. GO Cellular Component - Seed Genes Interactome. The first 10 most significantly enriched GO Cellular Component by genes involved in Seed Gene Interactome. The color of the bars corresponds to the p-value (the ligher, the more significant).

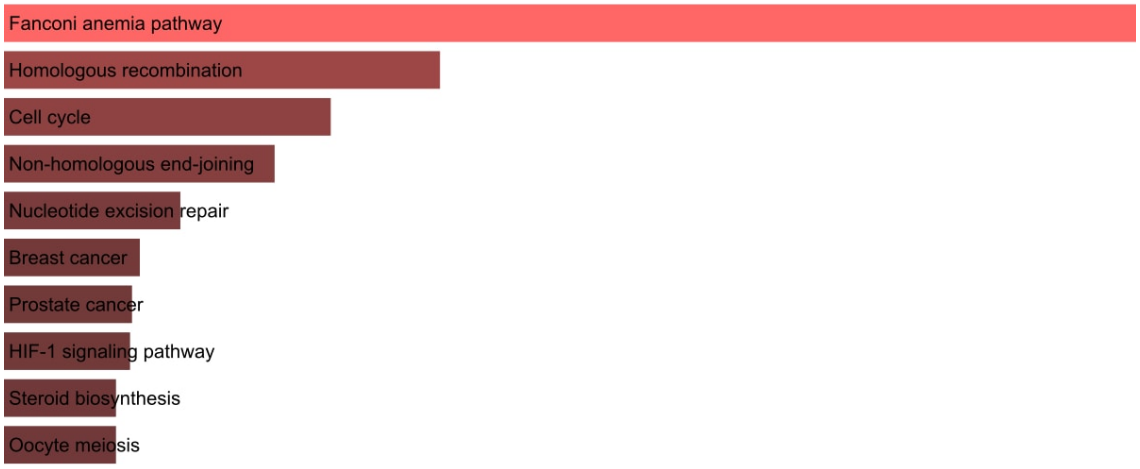

Figure 4. KEGG 2019 Human Pathways- Seed Genes Interactome. The first 10 most significantly enriched KEGG 2019 Human by genes involved in Seed Gene Interactome. The color of the bars corresponds to the p-value (the ligher, the more significant).

We noticed that collected seed genes are significantly involved in the biological processes of DNA repair and DNA metabolic processes. Also regarding molecular functions the most significant result corresponds to DNA binding. We obtained interesting result for pathways, as the most enriched pathway is The Fanconi anemia pathway. We checked that this pathway is required for the efficient repair of damaged DNA [1]. Mutations of such genes could have an impact, for instance on the mitotic spindle orientation or on the signalling responses because of the damaged DNA, and further leading to the primary microcephaly disease.

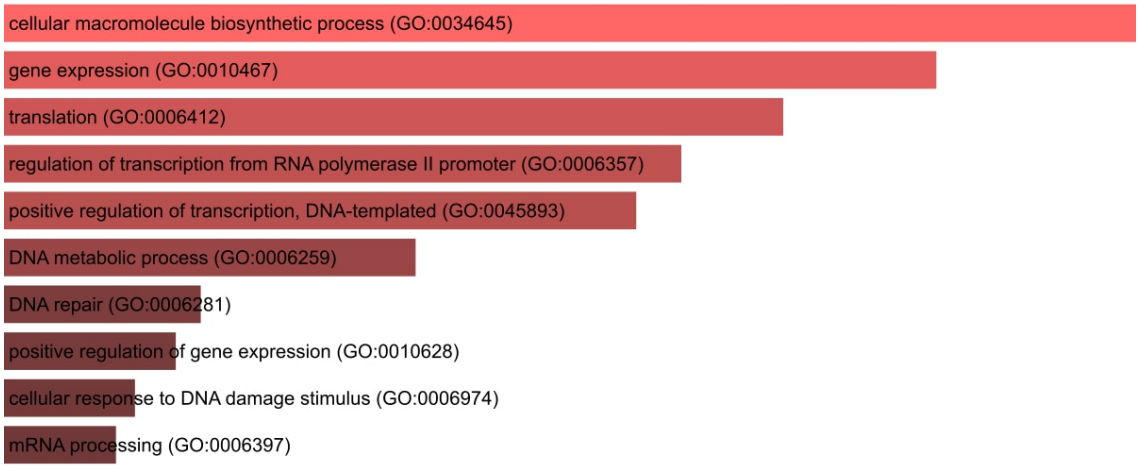

Figure 5. GO Biological Process - Union Interactome The first 10 most significantly enriched GO Biological Processes by genes involved in Union Interactome. The color of the bars corresponds to the p-value (the ligher, the more significant).

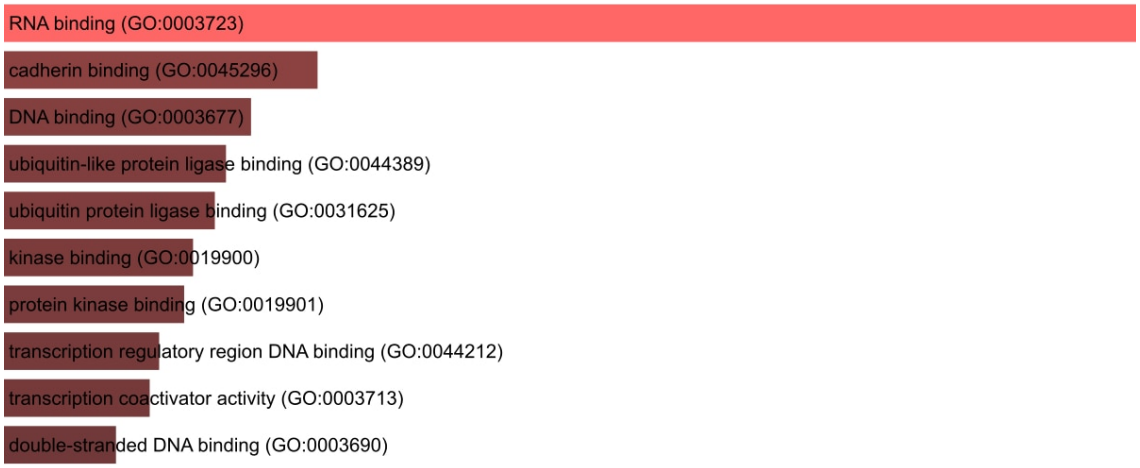

Figure 6. GO Molecular Function - Union Interactome The first 10 most significantly enriched GO Molecular Function by genes involved in Union Interactome. The color of the bars corresponds to the p-value (the ligher, the more significant).

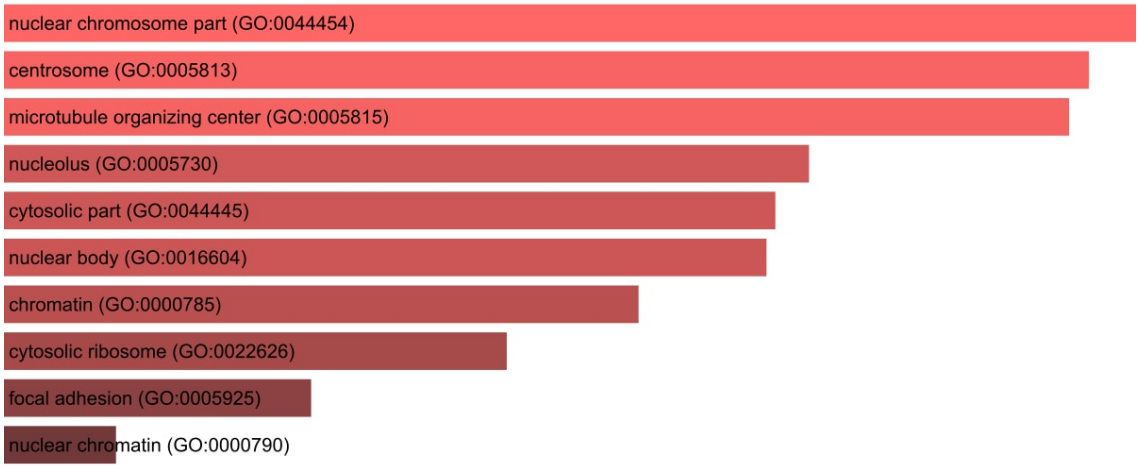

Figure 7. GO Cellular Component - Union Interactome The first 10 most significantly enriched GO Cellular Component by genes involved in Union Interactome. The color of the bars corresponds to the p-value (the ligher, the more significant).

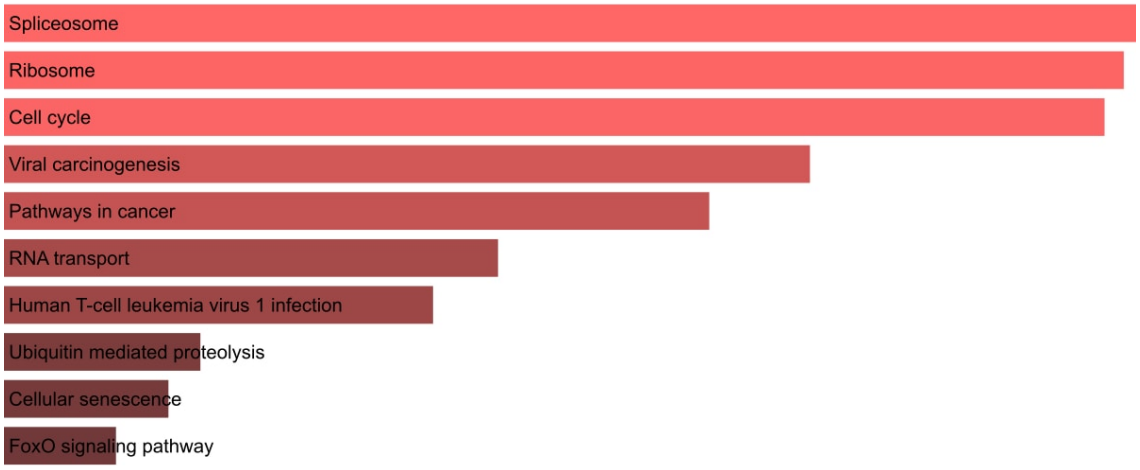

Figure 8. KEGG 2019 Human Pathways- Union Interactome he first 10 most significantly enriched KEGG 2019 Human by genes involved in Union Interactome. The color of the bars corresponds to the p-value (the ligher, the more significant).

Genes involved in the Union Interactome are significantly enriched in cellular macromolecule formation and gene expression biological processes. The results of molecular function enrichment analysis prove the role of these genes in RNA binding. Regarding the KEGG Pathways, the union interactome genes are involved in Spliceosome, Ribosome and Cell Cycle pathways. Both the spliceosomes and ribosomes contain central RNA components that give biomolecular complexes sequence-specific nucleic acid recognition, and comprehensive structural dynamics.

5. Network Measures

From the previously generated interactomes (SGI, I and U), we built the networks using Cytoscape. Graphical representation of the SGI network from Cytoscape is in the Figure 21 in Appendix.

Network Analyzer feature of Cytoscape enabled us to calculate Global Network Measures for each of the networks.

Table 4: **Global Network Measures for Seed Genes Interactome (SGI), Intersection Interactome (I) and Union Interactome (U).** Statistics are taken from the Network Analyser from Cytoscape

Interactome	**SGI**	**I**	**U**
Number of nodes	88	3609	5094
Number of links	400	52216	250489
Number of connected components	4	4	1
Number of isolated nodes	3	3	0
Average path length	3.392	2.807	2.376
Average degree	5.136	28.349	75.027
Average clustering coefficient	0.334	0.125	0.177
Network diameter	9	6	5
Network radius	5	4	3
Centralization	0.116	0.239	0.269

Union Interactome (U) forms connected network, thus the Largest Connected Component (U-LCC) is equal to network itself and all the global measures presented in the Table 4 in the column "U" are applicable. On the other hand, Intersection Interactome (I) network has 8 connected components and global measures for I-LCC are slightly different.

Graphical representation of the I-LCC network from Cytoscape is in the Figure 22 in Appendix.

Table 5: **Global Network Measures for Largest Connected Components of Intersection Interactome (I-LCC) and Union Interactome (U-LCC).** Statistics are taken from the Network Analyser from Cytoscape

	I-LCC	**U-LCC**
Number of nodes	3606	5094
Number of links	51165	250489
Number of connected components	1	1
Number of isolated nodes	3	0
Average path length	2.807	2.376
Average degree	28.373	75.027
Average clustering coefficient	0.125	0.177
Network diameter	6	5
Network radius	4	3
Centralization	0.239	0.269

Table 6: **Local Network Measures for the top 20 nodes of the Largest Connected Component of Intersection Interactome (I-LCC) based on Betweenness Centrality measure.**

Gene Symbol	**Node Degree**	**Betweenness Centrality**	**Eigenvector Centralit**	**Closeness Centrality**	**Ratio Betweenness/Node Degree**
NTRK1	889	0.112446	0.211863	0.544562	0.000126

Table 6: **Local Network Measures for the top 20 nodes of the Largest Connected Component of Intersection Interactome (I-LCC) based on Betweenness Centrality measure.**

Gene Symbol	Node Degree	Betweenness Centrality	Eigenvector Centralit	Closeness Centrality	Ratio Betweenness/Node Degree
XPO1	516	0.051495	0.099316	0.490343	0.0001
EGFR	463	0.038961	0.107076	0.496215	0.000084
TP53	454	0.030559	0.141028	0.49601	0.000067
ELAVL1	291	0.030249	0.057048	0.470258	0.000104
TRIM25	354	0.030103	0.055291	0.466184	0.000085
UBC	353	0.021751	0.085471	0.479133	0.000062
MYC	347	0.019264	0.0928	0.484934	0.000056
CTNNB1	222	0.017787	0.050556	0.467029	0.00008
ESR1	390	0.017269	0.133046	0.483179	0.000044
HSP90AA1	264	0.015506	0.071573	0.467999	0.000059
APP	192	0.014328	0.033212	0.440063	0.000075
OBSL1	379	0.014061	0.125369	0.476411	0.000037
RAD21	191	0.013972	0.052015	0.424767	0.000073
CDK2	314	0.013005	0.103113	0.469523	0.000041
CUL3	293	0.01288	0.069939	0.464742	0.000044
CREBBP	192	0.012153	0.038188	0.444623	0.000063
MCM2	276	0.01157	0.089904	0.460526	0.000042
HDAC1	228	0.011476	0.050123	0.449445	0.00005
NPM1	293	0.011472	0.113959	0.472044	0.000039

Table 7: **Local Network Measures for the top 20 nodes of the Largest Connected Component of Union Interactome (U-LCC) based on Betweenness Centrality measure.**

Gene Symbol	Node Degree	Betweenness Centrality	Eigenvector Centrality	Closeness Centrality	Ratio Betweenness/Node Degree
VIRMA	1464	0.0437	0.0970	0.5785	0.000030
ESR2	1475	0.0333	0.0895	0.5720	0.000023
MYC	1596	0.0297	0.1143	0.5774	0.000019
NTRK1	1519	0.0267	0.1022	0.5731	0.000018
EFTUD2	1448	0.0262	0.1114	0.5658	0.000018
APP	1401	0.0250	0.0491	0.5377	0.000018
TRIM25	1535	0.0237	0.0797	0.5504	0.000015
ELAVL1	1171	0.0217	0.0572	0.5370	0.000019
CTNNB1	1061	0.0210	0.0687	0.5370	0.000020
LARP7	970	0.0154	0.0790	0.5347	0.000016
XPO1	928	0.0137	0.0582	0.5334	0.000015
RECQL4	1074	0.0133	0.0951	0.5468	0.000012
EGFR	996	0.0132	0.0625	0.5357	0.000013
TP53	1257	0.0130	0.0741	0.5394	0.000010
JUN	1135	0.0129	0.0955	0.5489	0.000011
EGLN3	927	0.0129	0.0488	0.5343	0.000014
BRCA1	1664	0.0128	0.0795	0.5412	0.000008

Table 7: **Local Network Measures for the top 20 nodes of the Largest Connected Component of Union Interactome (U-LCC) based on Betweenness Centrality measure.**

Gene Symbol	Node Degree	Betweenness Centrality	Eigenvector Centrality	Closeness Centrality	Ratio Betweenness/Node Degree
HNRNPL	663	0.0117	0.0456	0.5217	0.000018
NR2C2	1004	0.0101	0.0881	0.5408	0.000010
CUL3	1355	0.0096	0.0865	0.5389	0.000007

6. Clustering methods for disease modules discovery

In order to understand the relationship between the organization of a network and its function, we conducted the cluster analysis of the protein interactome graphs (I-LCC and U-LCC). As the clustering method we applied the Markov Cluster (MLC) algorithm, implemented inside the ClusterMaker tool - a Cytoscape plugin with different clustering techniques.

The MCL algorithm simulates a flow/random walks on the graph by calculating subsequent powers of the associated adjacency matrix. With each iteration, an inflation step is used to enhance the contrast between regions of strong or weak flow in the graph. The process converges towards a division of the graph, with a set of high-flow areas (clusters) separated by no flow boundaries. The value of the inflation parameter strongly affects the number of clusters [**?**].

We applied hypergeometric test on clusters in order to find ones that contain a significant number of seed genes. Here we tested null Hypothesis H_0 that seed genes are not statistically overrepresented in a specific cluster against alternative hypothesis, H_1 that they are statistically overrepresented. We selected modules with more than 10 nodes and with a $p - value \leq 0.05$ and we consider those modules as putative disease modules.

In order to test how setting of inflation parameter affect clustering and putative disease discovery, we performed analysis with multiple values of inflation parameters. We summarized results for U-LCC and I-LCC in the tables below.

Table 8: **Putative Disease Modules found in the Largest Connected Component of Union Interactome (U-LCC) obtained for different values of inflation parameter.**

Inflation parameter	Number of seed genes	Total number of genes	Ratio = Seed/Total	p-value
2	6	81	0.074	6.18e-03
	2	17	0.118	4.68e-02
	6	15	0.4	2.86e-07
2.5	2	14	0.143	0.032
3	2	10	0.2	0.017
3.5	0	0	0	0

Table 9: **Putative Disease Modules found in the Largest Connected Component of Intersection Interactome (I-LCC) obtained for different values of inflation parameter.**

Inflation parameter	Number of seed genes	Total number of genes	Ratio = Seed/Total	p-value
2	6	22	0.273	0.000024
2.5	5	16	0.313	0.000058
3	4	13	0.308	0.000367
3.5	3	12	0.25	0.004117

Finally, we decided to keep the results obtained with inflation value 2.5.

Table 10: **Summary of Putative Disease Modules found in the Largest Conected Components of Union (U-LCC) and Intersection (I-LCC) Interactomes for inflation value = 2.5.**

	Source	
	U-LCC	I-LCC
Number of Seed Genes	2	5
Total Number of Genes	14	16
Ratio Seed/Total	0.143	0.313
p-value	0.032	0.000058

Union Interactome - the Largest Connected Component

We performed enrichment analysis of the genes from putative disease module obtained for U-LCC for GO categories and KEGG pathways. The following Figures 9, 10, 11, 12 corresponds to obtained charts with over-represented GO Biological Process, GO Molecular Function, GO Cellular Component and KEGG Pathways, respectively.

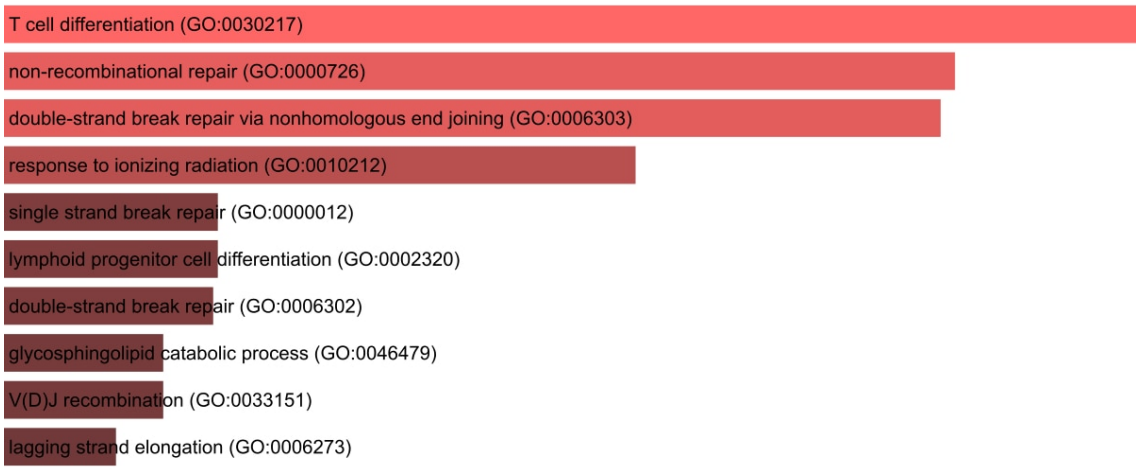

Figure 9. GO Biological Process - U-LCC Putative Disease Module The first 10 most significantly enriched GO Biological Processes by genes involved in U-LCC Putative Disease Module. The color of the bars corresponds to the p-value (the ligher, the more significant).

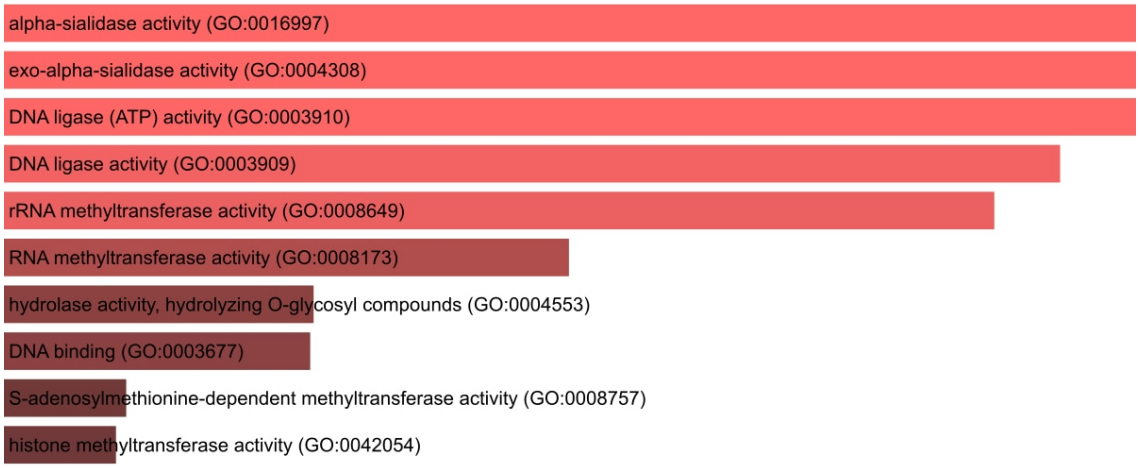

Figure 10. GO Molecular Function - U-LCC Putative Disease Module The first 10 most significantly enriched GO Molecular Function by genes involved in U-LCC Putative Disease Module. The color of the bars corresponds to the p-value (the ligher, the more significant).

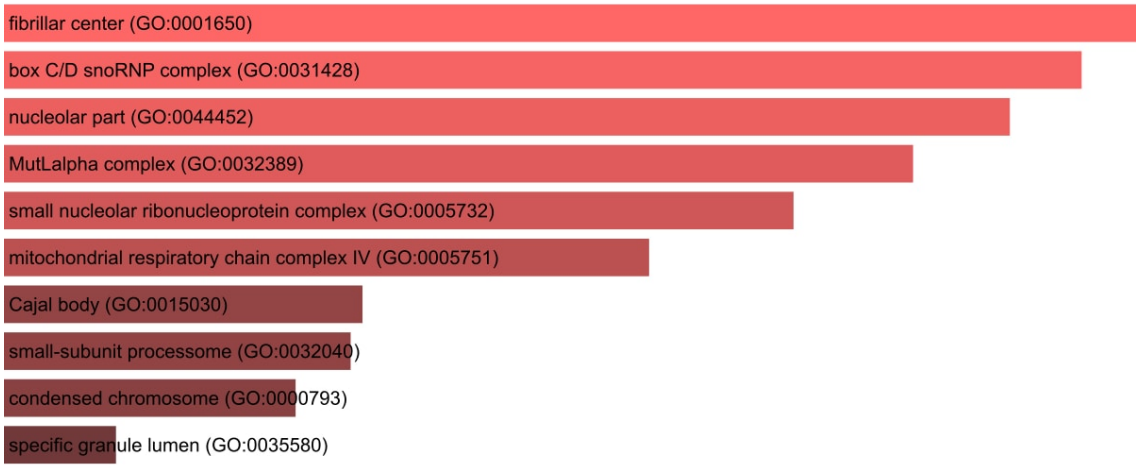

Figure 11. GO Cellular Component - U-LCC Putative Disease Module The first 10 most significantly enriched GO Cellular Component by genes involved in U-LCC Putative Disease Module. The color of the bars corresponds to the p-value (the ligher, the more significant).

Figure 12. KEGG 2019 Human Pathways- U-LCC Putative Disease Module The first 10 most significantly enriched KEGG 2019 Human by genes involved in U-LCC Putative Disease Module. The color of the bars corresponds to the p-value (the ligher, the more significant).

Intersection Interactome - the Largest Connected Component

We performed enrichment analysis of the genes from putative disease module obtained for I-LCC for GO categories and KEGG pathways. The following Figures 13, 14, 15, 16 corresponds to obtained charts with overrepresented GO Biological Process, GO Molecular Function, GO Cellular Component and KEGG Pathways, respectively.

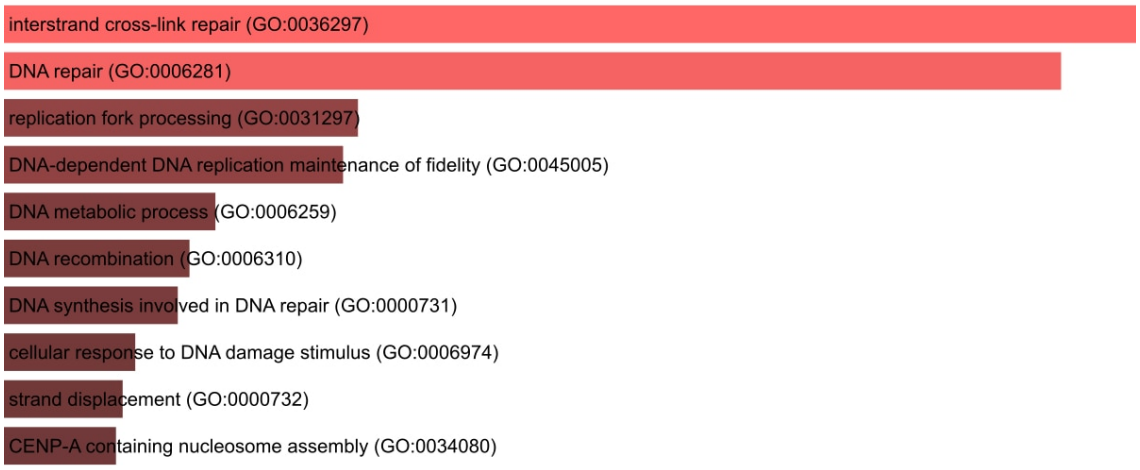

Figure 13. GO Biological Process - I-LCC Putative Disease Module The first 10 most significantly enriched GO Biological Process by genes involved in I-LCC Putative Disease Module. The color of the bars corresponds to the p-value (the ligher, the more significant).

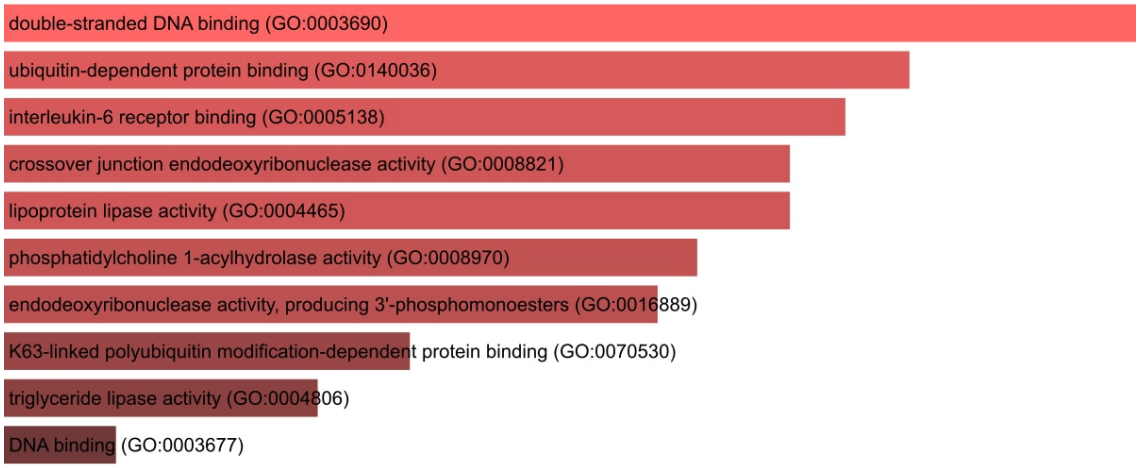

Figure 14. GO Molecular Function - I-LCC Putative Disease Module The first 10 most significantly enriched GO Molecular Function by genes involved in I-LCC Putative Disease Module. The color of the bars corresponds to the p-value (the ligher, the more significant).

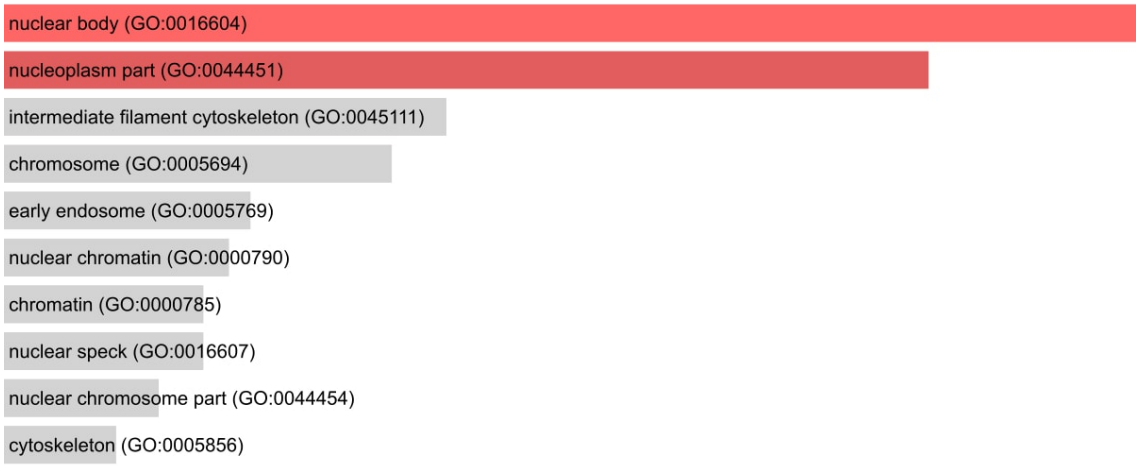

Figure 15. GO Cellular Component - I-LCC Putative Disease Module The first 10 most significantly enriched GO Cellular Component by genes involved in I-LCC Putative Disease Module. The color of the bars corresponds to the p-value (the ligher, the more significant).

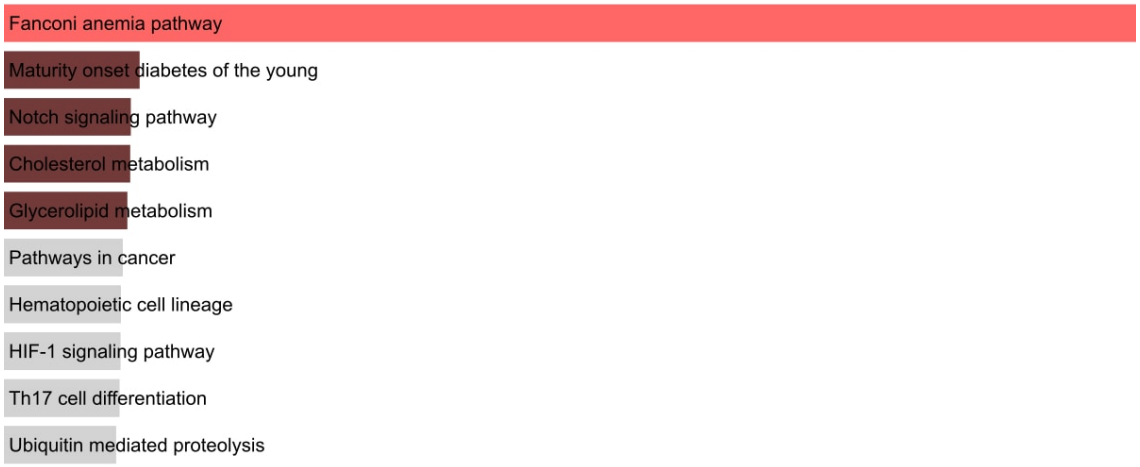

Figure 16. KEGG 2019 Human Pathways- I-LCC Putative Disease Module The first 10 most significantly enriched KEGG 2019 Human pathways by genes involved in I-LCC Putative Disease Module. The color of the bars corresponds to the p-value (the ligher, the more significant).

7. Putative disease genes by DIAMOnD

We computed the putative disease gene list with the use of DIAMOnD tool [8]. As the reference interactome we used the whole BioGrid interactome already used to collect PPIs. We limited the number of putative disease proteins to 200, as was suggested. The representation of first 30 genes is located in a Table 11.

Table 11: **First 30 Putative Disease Genes from Diamond**.

ID	Gene Symbol	ID	Gene Symbol	ID	Gene Symbol
1	BRCA1	**11**	C19orf40	**21**	MSH6
2	PCNA	**12**	C1orf86	**22**	TOPBP1
3	POLN	**13**	ETAA1	**23**	XRCC5
4	RPA1	**14**	RMI1	**24**	WRN
5	ATR	**15**	TOP3A	**25**	RPA3
6	CCNA2	**16**	RMI2	**26**	RFC1
7	RPA2	**17**	ATM	**27**	RAD50
8	APITD1	**18**	RAD51	**28**	MDC1
9	STRA13	**19**	MSH2	**29**	PRKDC
10	C17orf70	**20**	SPTA1	**30**	CDC5L

We performed enrichment analysis of the newly collected 200 genes for GO categories and KEGG pathways. The following Figures 17, 18, 19, 20 corresponds to obtained charts with overrepresented GO Biological Process, GO Molecular Function, GO Cellular Component and KEGG Pathways, respectively.

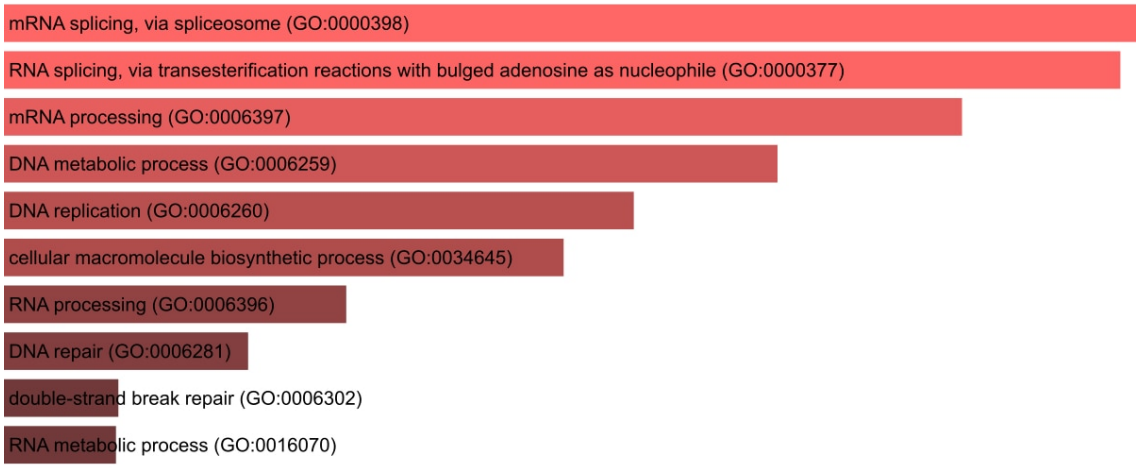

Figure 17. GO Biological Process - the putative disease protein list from DIAMOnD. The first 10 most significantly enriched GO Biological processes by genes involved in the Putative Disease Module. The color of the bars corresponds to the p-value (the ligher, the more significant)

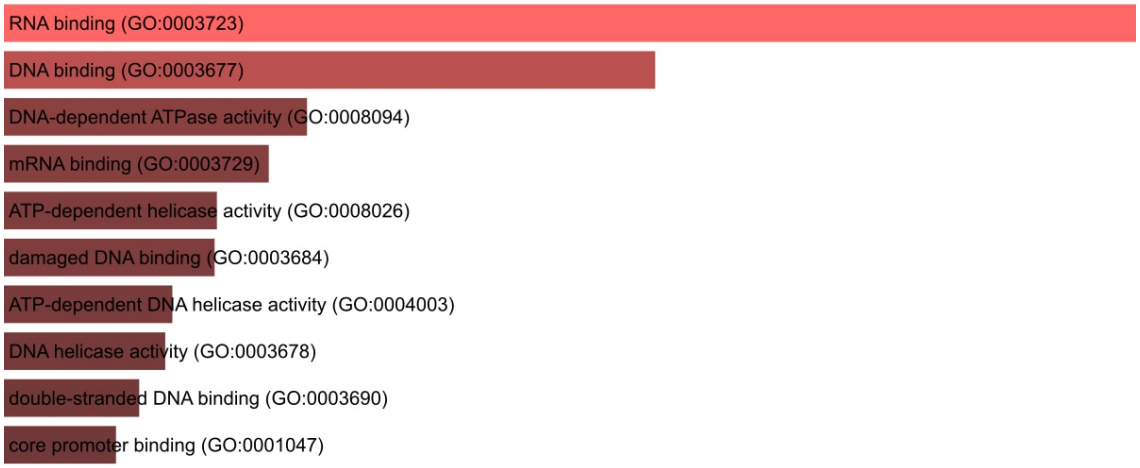

Figure 18. GO Molecular Function - the putative disease protein list from DIAMOnD. The first 10 most significantly enriched GO Molecular Function by genes involved in the Putative Disease Module. The color of the bars corresponds to the p-value (the ligher, the more significant)

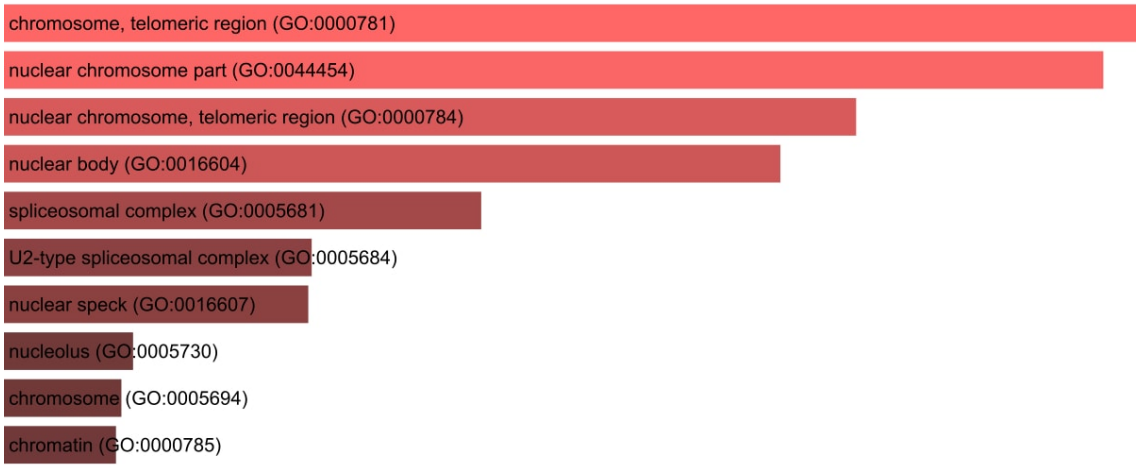

Figure 19. GO Cellular Component- the putative disease protein list from DIAMOnD. The first 10 most significantly enriched GO Cellular Component by genes involved in the Putative Disease Module. The color of the bars corresponds to the p-value (the ligher, the more significant)

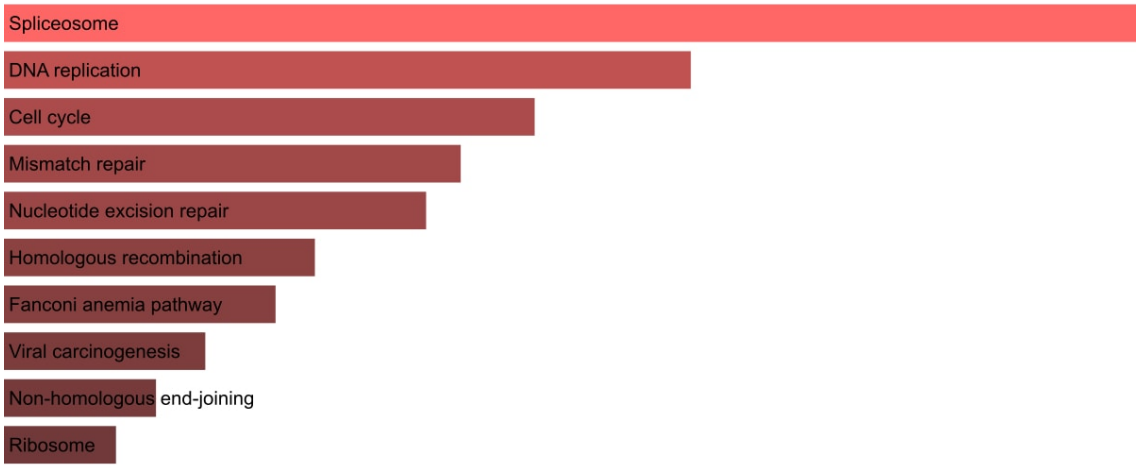

Figure 20. KEGG 2019 Human Pathways- the putative disease protein list from DIAMOnD. The first 10 most significantly enriched KEGG Pathways by genes involved in the Putative Disease Module. The color of the bars corresponds to the p-value (the ligher, the more significant)

8. Conclusions

We obtained 108 seed genes associated to Primary microcephaly disease. We found protein-protein interactions for 105 seed genes from Biogrid database and for 106 seed genes from IID database. The enrichment analysis of seed genes for gene ontologies and pathways confirmed that they are critical for many processes like, for instance the repair of damaged RNA and mutations of them could lead to cortical disorders like microcephaly. During the analysis, we were also able to find Putatice disease gene modules, a deeper analysis of which would help in further understanding the disease.

9. Appendix

Table 12: **Seed genes associated to Primary microcephaly disease from DisGeNet database.**

Entrez Gene ID	Gene Symbol	Uniprot AC	Protein name
546	ATRX	P46100	Transcriptional regulator ATRX
641	BLM	P54132	Bloom syndrome protein
675	BRCA2	P51587	Breast cancer type 2 susceptibility protein
990	CDC6	Q99741	Cell division control protein 6 homolog
1021	CDK6	Q00534	Cyclin-dependent kinase 6
1062	CENPE	Q02224	Centromere-associated protein E
1063	CENPF	P49454	Centromere protein F
1161	ERCC8	Q13216	DNA excision repair protein ERCC-8
1387	CREBBP	Q92793	CREB-binding protein
1499	CTNNB1	P35222	Catenin beta-1
1663	DDX11	Q96FC9	ATP-dependent DNA helicase DDX11
1717	DHCR7	Q9UBM7	7-dehydrocholesterol reductase
1729	DIAPH1	O60610	Protein diaphanous homolog 1
1763	DNA2	P51530	DNA replication ATP-dependent helicase/nuclease DNA2
1804	DPP6	P42658	Dipeptidyl aminopeptidase-like protein 6
1859	DYRK1A	Q13627	Dual specificity tyrosine-phosphorylation-regulated kinase 1A
1911	PHC1	P78364	Polyhomeotic-like protein 1
2072	ERCC4	Q92889	DNA repair endonuclease XPF
2073	ERCC5	P28715	DNA repair protein complementing XP-G cells
2074	ERCC6	P0DP91	Chimeric ERCC6-PGBD3 protein
2175	FANCA	O15360	Fanconi anemia group A protein
2176	FANCC	Q00597	Fanconi anemia group C protein
2177	FANCD2	Q9BXW9	Fanconi anemia group D2 protein
2178	FANCE	Q9HB96	Fanconi anemia group E protein
2187	FANCB	Q8NB91	Fanconi anemia group B protein
2188	FANCF	Q9NPI8	Fanconi anemia group F protein
2189	FANCG	O15287	Fanconi anemia group G protein
3376	IARS	P41252	Isoleucine–tRNA ligase, cytoplasmic
3479	IGF1	P05019	Insulin-like growth factor I
3480	IGF1R	P08069	Insulin-like growth factor 1 receptor
3832	KIF11	P52732	Kinesin-like protein KIF11
3981	LIG4	P49917	DNA ligase 4
4361	MRE11	P49959	Double-strand break repair protein MRE11
4613	MYCN	P04198	N-myc proto-oncogene protein
4683	NBN	O60934	Nibrin
4998	ORC1	Q13415	Origin recognition complex subunit 1
5000	ORC4	O43929	Origin recognition complex subunit 4
5116	PCNT	O95613	Pericentrin

Table 12: **Seed genes associated to Primary microcephaly disease from DisGeNet database.**

Entrez Gene ID	Gene Symbol	Uniprot AC	Protein name
5160	PDHA1	P08559	Pyruvate dehydrogenase E1 component subunit alpha, somatic form, mitochondrial
5859	QARS	P47897	Glutamine–tRNA ligase
5885	RAD21	O60216	Double-strand-break repair protein rad21 homolog
5889	RAD51C	O43502	DNA repair protein RAD51 homolog 3
5932	RBBP8	Q99708	DNA endonuclease RBBP8
6134	RPL10	P27635	60S ribosomal protein L10
6307	MSMO1	Q15800	Methylsterol monooxygenase 1
6491	STIL	Q15468	SCL-interrupting locus protein
6884	TAF13	Q15543	Transcription initiation factor TFIID subunit 13
7518	XRCC4	Q13426	DNA repair protein XRCC4
8243	SMC1A	Q14683	Structural maintenance of chromosomes protein 1A
8320	EOMES	O95936	Eomesodermin homolog
8573	CASK	O14936	Peripheral plasma membrane protein CASK
9126	SMC3	Q9UQE7	Structural maintenance of chromosomes protein 3
9343	EFTUD2	Q15029	116 kDa U5 small nuclear ribonucleoprotein component
9373	PLAA	Q9Y263	Phospholipase A-2-activating protein
9419	CRIPT	Q9P021	Cysteine-rich PDZ-binding protein
9662	CEP135	Q66GS9	Centrosomal protein of 135 kDa
9839	ZEB2	O60315	Zinc finger E-box-binding homeobox 2
10084	PQBP1	O60828	Polyglutamine-binding protein 1
10293	TRAIP	Q9BWF2	E3 ubiquitin-protein ligase TRAIP
10426	TUBGCP3	Q96CW5	Gamma-tubulin complex component 3
10479	SLC9A6	Q92581	Sodium/hydrogen exchanger 6
10617	STAMBP	O95630	STAM-binding protein
10733	PLK4	O00444	Serine/threonine-protein kinase PLK4
11284	PNKP	Q96T60	Bifunctional polynucleotide phosphatase/kinase
22995	CEP152	O94986	Centrosomal protein of 152 kDa
23001	WDFY3	Q8IZQ1	WD repeat and FYVE domain-containing protein 3
23141	ANKLE2	Q86XL3	Ankyrin repeat and LEM domain-containing protein 2

Table 12: **Seed genes associated to Primary microcephaly disease from DisGeNet database.**

Entrez Gene ID	Gene Symbol	Uniprot AC	Protein name
23594	ORC6	Q9Y5N6	Origin recognition complex subunit 6
25836	NIPBL	Q6KC79	Nipped-B-like protein
25886	POC1A	Q8NBT0	POC1 centriolar protein homolog A
25914	RTTN	Q86VV8	Rotatin
27229	TUBGCP4	Q9UGJ1	Gamma-tubulin complex component 4
29980	DONSON	Q9NYP3	Protein downstream neighbor of Son
51053	GMNN	O75496	Geminin
51124	IER3IP1	Q9Y5U9	Immediate early response 3-interacting protein 1
51199	NIN	Q8N4C6	Ninein
51574	LARP7	Q4G0J3	La-related protein 7
54820	NDE1	Q9NXR1	Nuclear distribution protein nudE homolog 1
55120	FANCL	Q9NW38	E3 ubiquitin-protein ligase FANCL
55215	FANCI	Q9NVI1	Fanconi anemia group I protein
55755	CDK5RAP2	Q96SN8	CDK5 regulatory subunit-associated protein 2
55835	CENPJ	Q9HC77	Centromere protein J
55869	HDAC8	Q9BY41	Histone deacetylase 8
57082	KNL1	Q8NG31	Kinetochore scaffold 1
57697	FANCM	Q8IYD8	Fanconi anemia group M protein
58497	PRUNE1	Q86TP1	Exopolyphosphatase PRUNE1
60386	SLC25A19	Q9HC21	Mitochondrial thiamine pyrophosphate carrier
63925	ZNF335	Q9H4Z2	Zinc finger protein 335
79648	MCPH1	Q8NEM0	Microcephalin
79728	PALB2	Q86YC2	Partner and localizer of BRCA2
79840	NHEJ1	Q9H9Q4	Non-homologous end-joining factor 1
80254	CEP63	Q96MT8	Centrosomal protein of 63 kDa
81620	CDT1	Q9H211	DNA replication factor Cdt1
83990	BRIP1	Q9BX63	Fanconi anemia group J protein
84126	ATRIP	Q8WXE1	ATR-interacting protein
84464	SLX4	Q8IY92	Structure-specific endonuclease subunit SLX4
84879	MFSD2A	Q8NA29	Sodium-dependent lysophosphatidylcholine symporter 1
84942	WDR73	Q6P4I2	WD repeat-containing protein 73
85378	TUBGCP6	Q96RT7	Gamma-tubulin complex component 6
93587	TRMT10A	Q8TBZ6	tRNA methyltransferase 10 homolog A
150468	CKAP2L	Q8IYA6	Cytoskeleton-associated protein 2-like
163786	SASS6	Q6UVJ0	Spindle assembly abnormal protein 6 homolog

Table 12: **Seed genes associated to Primary microcephaly disease from DisGeNet database.**

Entrez Gene ID	Gene Symbol	Uniprot AC	Protein name
259266	ASPM	Q8IZT6	Abnormal spindle-like microcephaly-associated protein
284403	WDR62	O43379	WD repeat-containing protein 62
286053	NSMCE2	Q96MF7	E3 SUMO-protein ligase NSE2
392636	AGMO	Q6ZNB7	Alkylglycerol monooxygenase
84919	PPP1R15B	Q5SWA1	protein phosphatase 1 regulatory subunit 15B

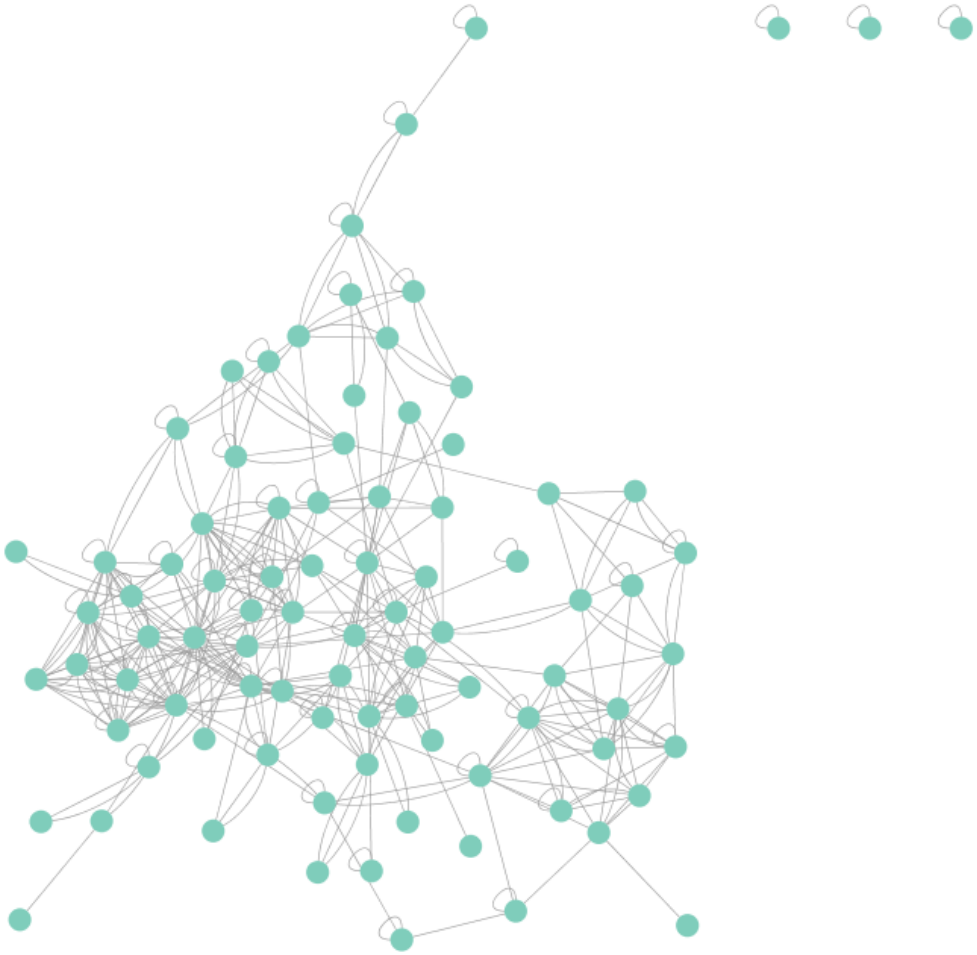

Figure 21. The figure of Seed Genes Interactome (SGI) Network from Cytoscape

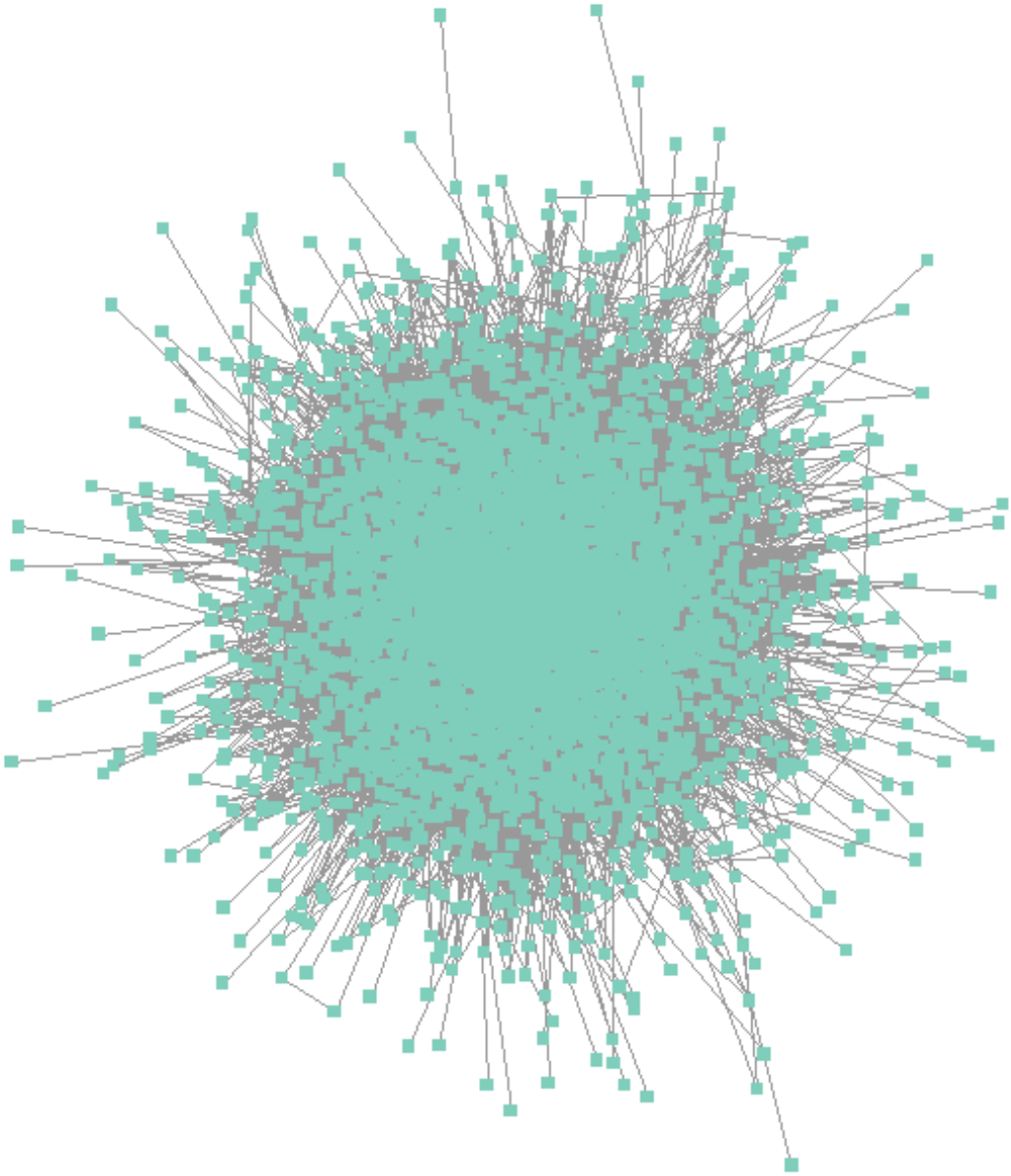

Figure 22. The figure of the Largest Connected Component Intersection Interactome (I-LCC) Network from Cytoscape

References

1 https://www.genome.jp/kegg-bin/show_pathway?map=ko03460&show_description=show
2 Finlay BL, Darlington RB. Linked regularities in the development and evolution of mammalian brains. Science, 1995;268(5217):1578-1584.
3 Faheem, M., Naseer, M.I., Rasool, M. et al. Molecular genetics of human primary microcephaly: an overview. BMC Med Genomics 8, S4 (2015) doi:10.1186/1755-8794-8-S1-S4
4 http://www.disgenet.org
5 https://www.cancer.gov/types/mesothelioma/hp
6 https://www.uniprot.org/
7 https://amp.pharm.mssm.edu/Enrichr/
8 https://github.com/barabasilab/DIAMOnD